Webology (ISSN: 1735-188X)
Volume 18, Number 6, 2021

An Efficient Pre-Processing Method Using Optimization Techniques For Heart Disease Prediction

V. Chezhiyan[1] , Dr. D.J. Evanjaline[2]

[1]Research Scholar, PG and Research Department of Computer Science, Rajah Serfoji Government College (Autonomous) (Affiliated to Bharathidasan University, Tiruchirappalli), Thanjavur, Tamilnadu, India.

[2]Assistant Professor, PG and Research Department of Computer Science, Rajah Serfoji Government College (Autonomous) (Affiliated to Bharathidasan University, Tiruchirappalli), Thanjavur, Tamilnadu, India.

ABSTRACT

The recent technology developments and innovations improves the life style of people through smart applications, sensors, wireless communication networks, etc., for all those technologies internet is the backbone and the information processing like accessing, distributing the necessary information is achieved through Internet of Things (IoT). IoT supports multi-disciplinary applications as an active entity in engineering, science and business discipline. Based on the user preference these applications and its services could be framed in IoT. Human daily activity recognition using mobile personal sensing technology plays a central role in the field of pervasive healthcare. Handling of huge volume of sensor data is a crucial issue in this domain for an effective decision-making system. In this research work, an effective pre-processing method is proposed using wrapper-based feature selection techniques. Harris Hawks Optimization (HHO) and Genetic Algorithm (GA) are hybridized to get the most pre-dominant features for the classification of heart disease where the data is obtained by the IoT wearable devices. The performance of the proposed pre-processing method is analysed with the existing feature selection techniques using different classifiers like Random Forest (RF), Gradient Boosting Tree (GBT) and Support Vector Machine (SVM) with various evaluation metrics like Accuracy, Precision, Recall and error rates.

KEYWORDS: Healthcare, Internet of Things (IoT), Optimization algorithms, Feature Selection, Classification

1. INTRODUCTION

Data is a familiar term among researchers as the research towards data management is still booming with inventive technologies [1] [2]. In the past few years, the amount of data is significantly raised due to the availability of services and ever-growing users. Vast amount of data is generated through sensors and actuators in real time environment which frames the Internet of Things (IoT). The data gathering architecture of IoT not only on sensors and also includes various sources like software applications, web resources etc., All these sources vast

amount of data and it requires a massive storage system. In addition to physical sensors, virtual sensors are recently developed which works based on a combination of data fusion from physical sensors which is used in the cloud environment. The collected information is termed as raw sensor data and it is collected, stored and processed as useful information which helps to solve data related necessities [3].

Modern electronic health records (EHR's) [4] are designed to capture and render clinical data from IoT during the health care process. Using them, health care providers can enter and access clinical data when it is needed. Through the presence of digital data, EHR's can incorporate decision support technologies to assist clinicians in providing better care. When adequate data is recorded in an EHR, data mining technologies can be used to automatically extract useful models and can assist in constructing the logic for decision support systems [5][6]. However, because the main function of EHR's is to store and report clinical data collected for the purpose of health care delivery, the characteristics of this data may not be optimal for data mining and other data analysis operations [7].

Through this research paper, an efficient pre-processing method is proposed which is used to remove the redundant and irrelevant data and feature from the raw dataset to improve the heart disease classification accuracy.

2. IMPORTANCE OF FEATURE SELECTION

Feature selection (FS) has attracted the attention of many researchers in the last few years due to the increasing sizes of datasets, which contain hundreds or thousands of columns (features). Typically, not all columns represent relevant values. Consequently, the noise or irrelevant columns could confuse the algorithms, leading to a weak performance of machine learning models. Different FS algorithms have been proposed to analyze highly dimensional datasets and determine their subsets of relevant features to overcome this problem. However, very often, FS algorithms are biased by the data. Thus, methods for ensemble feature selection (EFS) algorithms have become an alternative to integrate the advantages of single FS algorithms and compensate for their disadvantages.

Depending on the design of FS techniques [8][9][10], they are classified into three types of methods: filters, wrappers, and embedded. Each type defines advantages or disadvantages that are directly related to the context of the dataset. In general, these three types of FS techniques face typical problems, namely, (i) they have a good performance on a dataset, but by adding or removing instances, the performance decreases, (ii) they allow the removal of features quickly, but they are not capable of detecting redundant features, (iii) they need to have a correctly balanced dataset, and (iv) their performance is affected by the presence of noise in the data. Moreover, there is a large number of FS methods. However, there are no tools or solutions to determine objectively the algorithms, which would work best with the data of a particular domain.

3. RELATED WORKS

Haq, Amin Ul, et al [11] proposed a diagnosis system using machine learning methods for the detection of diabetes. The authors have proposed a filter method based on the Decision Tree (Iterative Dichotomiser 3) algorithm for highly important feature selection. Two ensemble

learning algorithms, Ada Boost and Random Forest, are also used for feature selection and we Li, Jian Ping, et al also compared the classifier performance with wrapper-based feature selection algorithms. Classifier Decision Tree has been used for the classification of healthy and diabetic subjects.

Zuo, Zheming, et al [12] aimed to reduce the number of features of EHR representation while improving the performance of the subsequent data analysis, e.g. classification. In this work, an efficient filter-based feature selection method, namely Curvature-based Feature Selection (CFS), is presented. The proposed CFS applied the concept of Menger Curvature to rank the weights of all features in the given data set.

Kogan, Emily, et al [13] The aim of this study was to use machine learning models to impute National Institutes of Health Stroke Scale (NIHSS) scores for all patients with newly diagnosed stroke from multi-institution electronic health record (EHR) data. NIHSS scores available in the Optum© de-identified Integrated Claims-Clinical dataset were extracted from physician notes by applying natural language processing (NLP) methods. Leveraging machine learning we identified the main factors in electronic health record data for assessing stroke severity, including death within the same month as stroke occurrence, length of hospital stay following stroke occurrence, aphagia/dysphagia diagnosis, hemiplegia diagnosis, and whether a patient was discharged to home or self-care.

Gronsbell, Jessica, et al [14] presented an automated feature selection method based entirely on unlabeled observations. The proposed method generates a comprehensive surrogate for the underlying phenotype with an unsupervised clustering of disease status based on several highly predictive features such as diagnosis codes and mentions of the disease in text fields available in the entire set of EHR data. A sparse regression model is then built with the estimated outcomes and remaining covariates to identify those features most informative of the phenotype of interest.

Awan, Saqib E., et al [15] The prediction of readmission or death after a hospital discharge for heart failure (HF) remains a major challenge. Modern healthcare systems, electronic health records, and machine learning (ML) techniques allow us to mine data to select the most significant variables (allowing for reduction in the number of variables) without compromising the performance of models used for prediction of readmission and death. Moreover, ML methods based on transformation of variables may potentially further improve the performance.

Spencer, Robinson, et al [16] experimentally assessed the performance of models derived by machine learning techniques by using relevant features chosen by various feature-selection methods. Four commonly used heart disease datasets have been evaluated using principal component analysis, Chi squared testing, ReliefF and symmetrical uncertainty to create distinctive feature sets. Then, a variety of classification algorithms have been used to create models that are then compared to seek the optimal features combinations, to improve the correct prediction of heart conditions.

Harerimana, Gaspard, et al [17] Traditional machine learning and statistical methods have failed to offer insights that can be used by physicians to treat patients as they need to obtain an expert opinion assisted features before building a benchmark task model. With the rise of deep learning methods, there is a need to understand how deep learning can save lives.

The purpose of this study was to offer an intuitive explanation for possible use cases of deep learning with EHR. The authors reflected on techniques that can be applied by health informatics professionals by giving technical intuitions and blue prints on how each clinical task can be approached by a deep learning algorithm.

Li, Jian Ping, et al [18] proposed novel fast conditional mutual information feature selection algorithm to solve feature selection problem. The features selection algorithms are used for features selection to increase the classification accuracy and reduce the execution time of classification system. Furthermore, the leave one subject out cross-validation method has been used for learning the best practices of model assessment and for hyperparameter tuning. The performance measuring metrics are used for assessment of the performances of the classifiers. The performances of the classifiers have been checked on the selected features as selected by features selection algorithms.

Hauser, Ronald G., et al [19] aimed to determine if machine learning models could predict CML using blood cell counts prior to diagnosis. The authors used 2 models (ie, XGBoost and LASSO) and 2 approaches to model selection to observe the effect of either choice on the study's results. Similar performance trends were observed between the 2 machine learning models and the 2 model selection approaches. Second, to control for variability in available laboratory data, we performed separate analyses on patients with complete and incomplete data and observed no significant difference in our results. Third, a patient was assigned to either test or train across all datasets, rather than assigning a patient to the test group in one dataset and the train group in another, eliminating an important source of variation.

Ali, Farman, et al. [20] proposed a smart healthcare system for heart disease prediction using ensemble deep learning and feature fusion approaches. First, the feature fusion method combines the extracted features from both sensor data and electronic medical records to generate valuable healthcare data. Second, the information gain technique eliminates irrelevant and redundant features, and selects the important ones, which decreases the computational burden and enhances the system performance. In addition, the conditional probability approach computes a specific feature weight for each class, which further improves system performance. Finally, the ensemble deep learning model is trained for heart disease prediction.

4. AN EFFICIENT PRE-PROCESSING METHOD USING OPTIMIZATION TECHNIQUES

4.1 Genetic Algorithm

In 1975, Holland proposed the GA based on the Darwin's theory of biological evolution [21]. Based on an initial population of chromosomes (i.e., solutions), this algorithm presents are flection of natural selection process where individuals of the highest fitness are selected for reproduction to generate the next-generation offsprings. The offsprings inherit characteristics of the parent and even transmit them to the proceeding generation. Should the parents exhibit a better fitness, the offsprings would exhibit a higher fitness as well, which is equal to a higher chance of survival compared to their parents. Figure 1 shows the flowchart of GA.

On this basis, in GA, the journey toward the optimal solution starts from an initial population of chromosomes that are initialized randomly. The size of this initial population is

linked to the nature and complexity of the problem at hand and remains unchanged throughout different iterations of this algorithm. The values assigned to each chromosme are analyzed by a fitness function. Thus, parent chromosomes are selected from the group of chromosomes with the highest fitness values compared to other chromosomes. This is realized by means of crossover and mutation operators. The crossover operator swaps, randomly, parts of one chromosome with those of another. The result is an offspring that inherits certain properties from each of the parent chromosomes rather than exactly resembling either of them. This operator sets the scene for achieving solutions of higher quality. When the mutation operator is applied to a chromosome, the value of one or more genes of a part of offspring chromosomes is changed randomly. As a next step, the newly generated solutions are assessed using the fitness function and the entire process is iterated until the stopping criterion is met, at which point the best solution is reported.

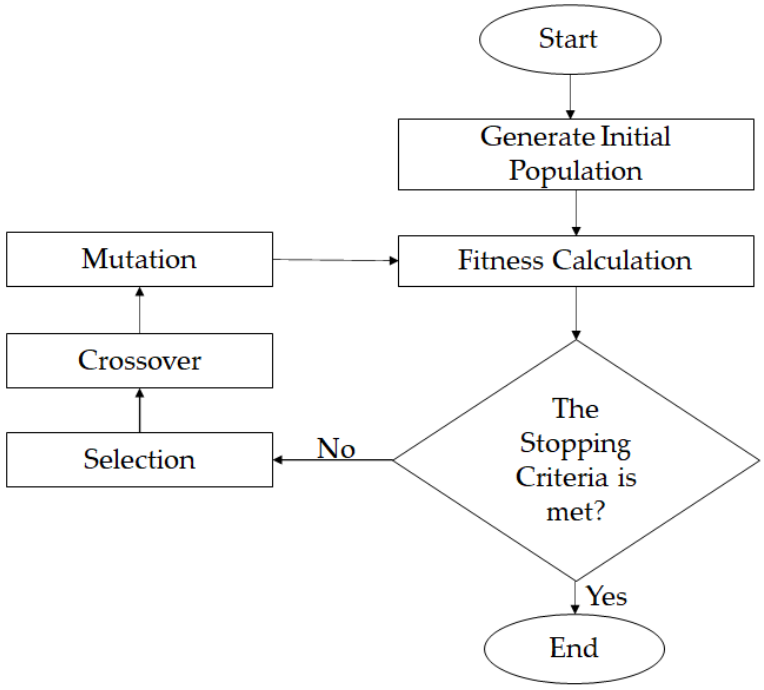

Figure 1: Flowchart of Genetic Algorithm

4.2 Harris Hawks Optimization (HHO)

HHO is a new meta-heuristic optimization algorithm introduced by Heidari et al. in 2019 [22]. HHO mimics the hunting mechanism of Harris Hawks in nature. The study of Harris hawks' behavior revealed that these birds use various sophisticated strategies in surprisingly attacking and hunting the fleeing prey (mostly a rabbit). As shown in the original publication of HHO, the mathematical modeling of this algorithm confirms its effectiveness in tackling diverse optimization problems. As any other population-based meta-heuristic optimizer, HHO generates a population of search agents and updates these search agents using exploration and exploitation phases. The exploration of this algorithm has two stages, while the exploitation consists of four stages.

4.2.1 Initialization Phase

In this phase, the objective function and the search-space are defined. Also, the initial population-based Genetic Algorithm are initiated. In addition, all parameter values are set.

4.2.2 Exploration Phase

In this phase, all Harris hawks are considered as candidate solutions. In each iteration, the fitness value is computed for all these possible solutions based on the intended prey. Two approaches are applied to mimic the exploration performances of Harris Hawks in the search space specified in (1):

$$X(t+1) = \begin{cases} X_{rand}(t) - r_1 \left| X_{rand}(t) - 2r_2 X(t) \right| & q \geq 0.5 \\ (X_{rabbit}(t) - X_m(t)) - r_3(LB + r_4(UB - LB)) & q < 0.5, \end{cases} \quad (1)$$

Where $X(t+1)$ is the position-of Hawks in second iteration t. $X_{rabbit}(t)$ is the prey position and the $X_{rand}(t)$ stands for the random solution chosen in the current population. $X(t)$ is the position vector of Hawks in the current iteration t, the r_1, r_2, r_3, r_4 and q are random scaled factor within [0; 1], which are updated in each iteration, LB and UB are the upper and lower bounds of variables, and the X_m is the average number of the solutions.

This intended approach generates the positions of Hawks within (UB - LB) bounds based on two rules; 1) create the solutions based on randomly selected hawk from the current population and the other hawks. 2) create the solutions based on the prey location, the average position of Hawks, and random scaled factors. While r_3 is a scaling factor, once the value of r_4 is close to 1, it will help increase the randomness of the rule. In this rule, a randomly scaled movement length is added to LB. A random scaled component is considered to provide more diversification techniques to explore different areas of the feature space. The average position of hawks (solutions) is formulated in (2).

$$X_m(t) = \frac{1}{N} \sum_{i=1}^{N} X_i(t) \quad (2)$$

Where $X_m(t)$ is the average number of the solutions in the current iteration. N indicates all possible solutions. $X_i(t)$ implies the location of each solution in iteration t, which created based chaos theory. Usually, in Eq. (1), rule one is applied when the hawk uses the information from the random hawks to catch the prey. While rule two is applied when all hawks share the best solution and the best hawk employed.

4.2.3 Transition from Exploration to Exploitation

This phase explains the movement of HHO from exploration to exploitation, based on the energy of the prey (E). HHO assumes that the energy of prey is reduced gradually through the escaping actions. E_o is the initial energy decreases from [1, -1], modeled in (3).

$$E = 2E_0 \left(1 - \frac{t}{T}\right), \quad E_0 \in [-1, 1] \quad (3)$$

where T indicates the maximum number of iterations, and t is the current iteration.

4.2.4 Exploitation Phase

In this phase, the exploitation phase is accomplished using four approaches at parameter sets. These approaches are based on the position identified in the exploration phase. However, the prey tries to escape frequently, while the hawks tracing and try to catch it. HHO exploitation is mimic the attacking strategy of the Hawks by using four possible approaches. These approaches are the soft besiege, hard besiege, soft besiege with progressive rapid dives, and hard besiege with progressive rapid dives. These approaches based on two variables r and |E|, which specify the executed approach. Where |E| is the escaping energy of the prey, r refers to the probability of escaping, where $r < 0.5$ indicates the higher possibility for the prey to escape successfully and $r \geq 0.5$ for unsuccessfully escape. A summary of these approaches are presented as follows:

In the soft besiege approach, where $r \geq 0.5$ and $|E| \geq 0.5$, the rabbit still has some energy to escape, while the hawks are softly encircling the prey make it lose more energy before performing the surprise pounce. Soft besiege mathematically formulated in (4), (5), and (6).

$$X(t+1) = \Delta X(t) - E\,|JX_{rabbit} - X(t)| \qquad (4)$$

$$\Delta X(t) = X_{rabbit} - X(t) \qquad (5)$$

$$J = 2(1 - r_5), \quad r_5 \in [0, 1] \qquad (6)$$

Where Δx(t) is the difference among the position vector of the prey and the current location in iteration t, and J presents the jump power of the prey and r_5 is a random variable.

In the hard besiege strategy, where $r \geq 0.5$ and $|E| < 0.5$, the prey is tired with a weak escaping chance. In this condition, the hawk hardly encircles the prey to perform the final surprise pounce. Thus, the solution is updated using (7).

$$X(t+1) = X_{rabbit}(t) - E\,|\Delta X(t)| \qquad (7)$$

Eq. (8) shows the soft besiege with progressive rapid dives approach. In this condition $r < 0.5$ and $|E| \geq 0.5$, the prey still has the energy to escape. The hawk moves smartly around the prey and patiently dives before the surprised pounce. This action is considered as intelligent soft besiege, where the position of the hawks is updated in two steps. In the first step, the hawks move toward the prey by estimating the next move of the prey as formula (8):

$$Y = X_{rabbit}(t) - E|JX_{rabbit}(t) - X(t)| \qquad (8)$$

In the second step, the hawk decided whether to dive or not, based on the comparison between the previous dive and the possible result. If it is not, the hawks producing irregular dive, based on the Levy Flight (LF) concept, as formulated in (9):

$$Z = Y + S \times LF(Dim) \qquad (9)$$

where the dimension of solutions is defined as Dim, S is a random vector of size 1x dim. LF is the Levy Flight function calculated using (10):

$$LF(x) = 0.01 \times \frac{u \times \sigma}{|v|^{\frac{1}{\beta}}}, \quad \sigma = \left(\frac{\Gamma(1+\beta) \times \sin(\frac{\pi\beta}{2})}{\Gamma(\frac{1+\beta}{2}) \times \beta \times 2^{(\frac{\beta-1}{2})}} \right)^{\frac{1}{\beta}} \qquad (10)$$

Where β is a default constant set automatically to 1.5, and u, v are random values within [0,1]. Therefore, updating the Harris hawks positions in with progressive rapid dives can be formulated in (11):

$$X(t+1) = \begin{cases} Y & if\ F(Y) < F(X(t)) \\ Z & if\ F(Z) < F(X(t)) \end{cases} \qquad (11)$$

where Y and Z are performed using (8) and (9), and both refer to the new iteration's next location.

The last approach is called hard besiege with progressive rapid dives, where $r < 0.5$ and $|E| < 0.5$. In this condition, the prey has no energy to escape, and the Harris hawks attempt to reach the prey by rapid dives before performing a surprise pounce to catch the prey. The movement of the hawks in the condition is formulated in (12):

$$X(t+1) = \begin{cases} Y & if\ F(Y) < F(X(t)) \\ Z & if\ F(Z) < F(X(t)) \end{cases} \qquad (12)$$

where Y is set as in (13), and Z is updated as in (14).

$$Y = X_{rabbit}(t) - E\,|JX_{rabbit}(t) - X_m(t)| \qquad (13)$$

$$Z = Y + S \times LF(Dim) \qquad (14)$$

Finally, the classification accuracy computed using the fitness function set in Eq. (15). The fitness function includes the computation of classification error, as mathematically formulated in (15):

$$Fitness = \alpha\gamma_R(D) + \beta\frac{|R|}{|N|} \qquad (15)$$

Where αγR(D) refer to the classification error rate of the used classifiers. Besides, ⌈R⌉ is a cardinal number of the selected subset and ⌈N⌉ is a total number of features in the dataset α and β are two parameters corresponding to the importance of classification quality and subset length, $\alpha \in [0,1]$ and $\beta = (1-\alpha)$.

4.3 Proposed an Efficient Pre-Processing Method using GA and HHO techniques

In this study, feature selection is regarded as a multi-objective optimization problem, in which two contradictory goals must be achieved. These goals are to minimize the number of selected features and maximize the classification accuracy. In other words, to reach a minimum number of selected features in the solution that leads to higher classification accuracy. Every solution is calculated according to the proposed fitness function, which depends on the classifiers, to obtain the classification accuracy of the solution as well as the number of selected features. To balance the number of selected features in each solution (to be minimum) with the classification accuracy (to be maximum), we have chosen the fitness function in the equation (15) is applied for evaluating the search agents in the algorithm.

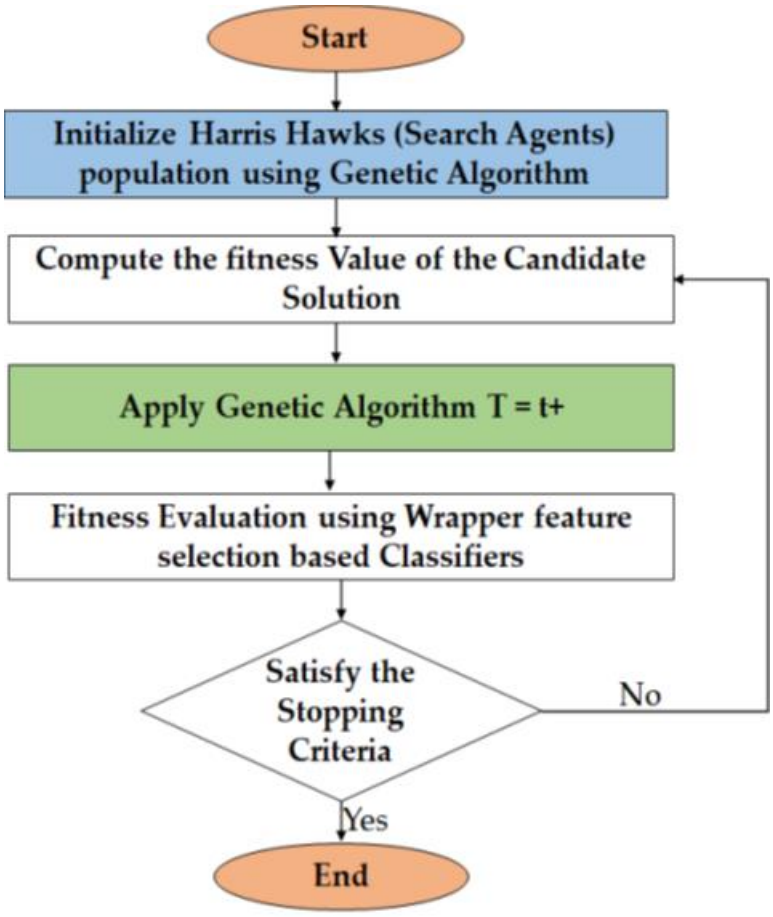

Figure 2: Flowchart of Proposed Genetic based Harris Hawks Feature Selection (GBHHFS) Method

Based on the previous studies, which utilized HHO for solving different problems and confirmed its outperformance in comparison to other recent and well-known optimization algorithms. However, the standard HHO algorithm suffers from two significant problems when applied to high- dimensional problems such as the feature selection problem. These problems are including 1) problem of solutions diversity; 2) problem of local optima. Therefore, to improve the HHO algorithm and make it suitable for the feature selection problem, two main improvements are introduced in this study to solve the weakness of the HHO algorithm. The improvement consists of using the GA algorithm with the HHO algorithm to enhance its exploitation and avoid being stuck in local optima.

The improvement is to embed the GA in the HHO algorithm to enhance its local searchability. This embedding will improve the exploitation capability of the algorithm. GA is used to improve the current best solution at the end of each HHO iteration.

Algorithm: An Efficient Pre-Processing Method using Optimization Techniques GBHHFS Method

Input: The population size N and maximum number of iterations T.
Output: The Location of a rabbit (feature subset) and its fitness value.

Step 1: Initialize the population $X_i (i = 1,2, \dots, N)$
Step 2: while (fitness value != stopping criteria) do.
Step 2.1: Compute the fitness values of hawks.
Step 2.2: Set X_{rabbit} as the location of rabbit (best location).
Step 2.3: for (each hawk (X_i)) **do**
Update the initial energy E_o and jump strength J
Step 3: $E_o = 2\text{rand}() - 1, j = 2(1 - \text{rand}())$
Step 4: Update the E using equation (3) [Exploration Phase]
Step 4.1: if $(|E|) \geq 1$ then
Step 4.2: Update the location vector using (1)
Step 4.3: if $(|E|) < 1$ then [Exploitation Phase]
If $(r \geq 0.5$ and $|E| \geq 0.5)$ then [Soft besiege]
Update the location vector using Eq. (4)
Elseif $(r \geq 0.5$ and $|E| < 0.5)$ then [Hard besiege]
Update the location vector using Eq. (7).
Elseif $(r < 0.5$ and $|E| \geq 0.5)$ then [Soft besiege with progressive rapid dives]
Update the location vector using Eq. (11)
Elseif $(r < 0.5$ and $|E| < 0.5)$ then [Hard besiege with progressive rapid dives]
Update the location vector using Eq. (12)
Step 5: Apply GA
Step 6: T = t+
Step 7: Return X_{rabbit}

5. RESULT AND DISCUSSION

5.1 Performance Metrics

The performance of the proposed Feature Selection method is evaluated with their existing Wrapper based feature selection methods like HHO, GA, Particle Swarm Optimization (PSO), Artificial Bee Colony (ABC) using classification techniques like Gradient Boosting Tree (GBT), Support Vector Machine and Random Forest. Table 1 depicts the performance metrics used to evaluate the performance of the existing and proposed feature selection methods for the given dataset. The dataset used in this research work is considered from the Kaggle repository [23].

Table 1: Performance Metrics

Metrics	Equation
Accuracy	$\frac{TP + TN}{TP + FN + TN + FP}$
True Positive Rate (TPR) (Sensitivity or Recall)	$\frac{TP}{TP + FN}$

False Positive Rate (FPR)	$\frac{FP}{FP + TN}$
Precision	$\frac{TP}{TP + FP}$
True Negative Rate (Specificity)	1- False Positive Rate (FPR)
Miss Rate	1-True Positive Rate (TPR)
False Discovery Rate	1- Precision

5.4 Performance Analysis of the Proposed GBHHFS method

Table 2 depicts the Classification Accuracy (in %) obtained for the heart disease dataset using original dataset, Proposed GBHHFS method, HHO, GA, ABC and PSO method processed datasets using GBT, RF and SVM. From the table 2, it is clear that the proposed GBHHFS method with GBT gives better accuracy than other existing feature selection methods.

Table 2: Classification Accuracy (in %) obtained for the Heart Disease dataset using original dataset, Proposed GBHHFS method, HHO, GA, ABC and PSO method processed datasets using GBT, RF and SVM

Feature Selection Methods	Classification Accuracy (in %) by Classification Techniques		
	GBT	RF	SVM
Original dataset	55.38	45.63	43.32
Proposed GBHHFS method	95.78	92.29	89.63
HHO	73.85	63.81	58.45
GA	74.99	71.86	68.02
ABC	69.74	65.95	63.54
PSO	68.68	62.65	61.45

Table 3 gives the True Positive Rate (in %) obtained for the heart disease dataset using original dataset, Proposed GBHHFS method, HHO, GA, ABC and PSO method processed datasets using GBT, RF and SVM. From the table 3, it is clear that the proposed GBHHFS method with GBT gives better TPR than other existing feature selection methods.

Table 3: True Positive Rate (in %) obtained for the heart disease dataset using original dataset, Proposed GBHHFS method, HHO, GA, ABC and PSO method processed datasets using GBT, RF and SVM

Feature Selection Methods	True Positive Rate (in %) by Classification Techniques		
	GBT	RF	SVM
Original dataset	54.49	44.54	42.23
Proposed GBHHFS method	95.59	91.38	89.72
HHO	75.81	72.95	69.13
GA	74.96	64.92	59.56
ABC	68.86	64.86	62.63

PSO	67.77	61.56	60.53

Table 4 gives the False Positive Rate (in %) obtained for the heart disease dataset using original dataset, Proposed GBHHFS method, HHO, GA, ABC and PSO method processed datasets using GBT, RF and SVM. From the table 4, it is clear that the proposed GBHHFS method with GBT gives reduced FPR than other existing feature selection methods.

Table 4: False Positive Rate (in %) obtained for the Heart Disease dataset using original dataset, Proposed GBHHFS method, HHO, GA, ABC and PSO method processed datasets using GBT, RF and SVM

Feature Selection Methods	False Positive Rate (in %) by Classification Techniques		
	GBT	**RF**	**SVM**
Original dataset	53.61	64.17	65.69
Proposed GBHHFS method	5.94	6.41	9.54
HHO	22.42	30.18	33.47
GA	27.53	33.62	34.47
ABC	38.82	44.51	45.84
PSO	41.72	47.34	48.73

Table 5 gives the Precision (in %) obtained for the Heart Disease dataset using original dataset, Proposed GBHHFS method, HHO, GA, ABC and PSO method processed datasets using GBT, RF and SVM. From the table 5, it is clear that the proposed GBHHFS method with GBT gives maximum Precision than other existing feature selection methods.

Table 5: Precision (in %) obtained for the heart disease dataset using original dataset, Proposed GBHHFS method, HHO, GA, ABC and PSO method processed datasets using GBT, RF and SVM

Feature Selection Methods	Precision (in %) by Classification Techniques		
	GBT	**RF**	**SVM**
Original dataset	66.81	53.92	46.76
Proposed GBHHFS method	96.52	90.53	80.66
HHO	79.25	71.38	62.74
GA	78.72	69.82	67.81
ABC	65.88	62.76	58.97
PSO	60.52	61.53	57.85

Table 6 gives the Specificity (in %) obtained for the heart disease dataset using original dataset, Proposed GBHHFS method, HHO, GA, ABC and PSO method processed datasets using GBT, RF and SVM. From the table 6, it is clear that the proposed GBHHFS method with GBT gives maximum specificity than other existing feature selection methods.

Table 6: Specificity (in %) obtained for the heart disease dataset using original dataset, Proposed GBHHFS method, HHO, GA, ABC and PSO method processed datasets using GBT, RF and SVM

Feature Selection Methods	Specificity (in %) by Classification Techniques		
	GBT	RF	SVM
Original dataset	46.39	35.83	34.31
Proposed GBHHFS method	94.06	93.59	90.46
HHO	77.58	69.82	66.53
GA	72.47	66.38	65.53
ABC	61.18	55.49	54.16
PSO	58.28	52.66	51.27

Table 7 gives the Miss Rate (in %) obtained for the heart disease dataset using original dataset, Proposed GBHHFS method, HHO, GA, ABC and PSO method processed datasets using GBT, RF and SVM. From the table 7, it is clear that the proposed GBHHFS method with GBT gives reduced miss rate than other existing feature selection methods.

Table 7: Miss Rate (in %) obtained for the heart disease dataset using original dataset, Proposed GBHHFS method, HHO, GA, ABC and PSO method processed datasets using GBT, RF and SVM

Feature Selection Methods	Classification Techniques		
	GBT	RF	SVM
Original dataset	45.51	55.46	57.77
Proposed GBHHFS method	4.41	8.62	10.28
HHO	24.19	27.05	30.87
GA	25.04	35.08	40.44
ABC	31.14	35.14	37.37
PSO	32.23	38.44	39.47

Table 8 gives the False Discovery Rate obtained by the original dataset, existing and Proposed feature selection processed datasets using RF, GBT and ANN classification techniques. From the table 8, it is clear that the proposed GBHHFS method with GBT gives reduced false discovery rate than other existing feature selection methods.

Table 8: False Discovery Rate (in %) obtained for the heart disease dataset using original dataset, Proposed GBHHFS method, HHO, GA, ABC and PSO method processed datasets using GBT, RF and SVM

Feature Selection Methods	Classification Techniques		
	GBT	RF	SVM
Original dataset	33.19	46.08	53.24
Proposed GBHHFS method	3.48	9.47	19.34
HHO	20.75	28.62	37.26

GA	21.28	30.18	32.19
ABC	34.12	37.24	41.03
PSO	39.48	38.47	42.15

6. CONCLUSION

Due to big data progress in biomedical and healthcare communities, accurate study of medical data benefits early disease recognition, patient care and community services. When the quality of medical data is incomplete the exactness of study is reduced. Moreover, different regions exhibit unique appearances of certain regional diseases, which may result in weakening the prediction of disease outbreaks. Through this research work, optimization techniques-based feature selection is proposed to enhance the disease prediction accuracy of the classification. HHO and GA is hybridized to extract the most pre-dominant features from the disease's datasets. The accuracy of the proposed Genetic Based Harris Hawks Feature Selection method is evaluated with various metrics using three different classifiers like GBT, SVM and RF. From the results obtained, it is shown that the proposed GBHHFS method gives better result with GBT classifier in terms of Accuracy, TPR, FPR, Specificity, Miss Rate, False discovery rate than other feature selection techniques and classifiers.

REFERENCES

[1] Grote, Thomas, and Philipp Berens. "On the ethics of algorithmic decision-making in healthcare." Journal of medical ethics 46.3 (2020): 205-211.

[2] Ahmad, Muhammad Aurangzeb, et al. "Fairness in machine learning for healthcare." Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020.

[3] Rose, Sherri. "Machine learning for prediction in electronic health data." JAMA network open 1.4 (2018): e181404-e181404.

[4] Wong, Jenna, et al. "Using machine learning to identify health outcomes from electronic health record data." Current epidemiology reports 5.4 (2018): 331-342.

[5] Kogan, Emily, et al. "Assessing stroke severity using electronic health record data: a machine learning approach." BMC medical informatics and decision making 20.1 (2020): 1-8.

[6] Masino, Aaron J., et al. "Machine learning models for early sepsis recognition in the neonatal intensive care unit using readily available electronic health record data." PloS one 14.2 (2019): e0212665.

[7] Gianfrancesco, Milena A., et al. "Potential biases in machine learning algorithms using electronic health record data." JAMA internal medicine 178.11 (2018): 1544-1547.

[8] Poornappriya, T. S., and M. Durairaj. "High relevancy low redundancy vague set- based feature selection method for telecom dataset." Journal of Intelligent & Fuzzy Systems Preprint: 1-18.

[9] Durairaj, M., and T. S. Poornappriya. "Choosing a spectacular Feature Selection technique for telecommunication industry using fuzzy TOPSIS MCDM." International Journal of Engineering & Technology 7.4 (2018): 5856-5861.

[10] Durairaj, M., and T. S. Poornappriya. "Why Feature Selection in Data Mining Is Prominent? A Survey." International Conference on Artificial Intelligence, Smart Grid and Smart City Applications. Springer, Cham, 2019

[11] Haq, Amin Ul, et al. "Intelligent machine learning approach for effective recognition of diabetes in E-healthcare using clinical data." Sensors 20.9 (2020): 2649.

[12] Zuo, Zheming, et al. "Curvature-based feature selection with application in classifying electronic health records." Technological Forecasting and Social Change 173 (2021): 121127.

[13] Kogan, Emily, et al. "Assessing stroke severity using electronic health record data: a machine learning approach." BMC medical informatics and decision making 20.1 (2020): 1-8.

[14] Gronsbell, Jessica, et al. "Automated feature selection of predictors in electronic medical records data." Biometrics 75.1 (2019): 268-277.

[15] Awan, Saqib E., et al. "Feature selection and transformation by machine learning reduce variable numbers and improve prediction for heart failure readmission or death." PloS one 14.6 (2019): e0218760.

[16] Spencer, Robinson, et al. "Exploring feature selection and classification methods for predicting heart disease." Digital health 6 (2020): 2055207620914777.

[17] Harerimana, Gaspard, et al. "Deep learning for electronic health records analytics." IEEE Access 7 (2019): 101245-101259.

[18] Li, Jian Ping, et al. "Heart disease identification method using machine learning classification in e-healthcare." IEEE Access 8 (2020): 107562-107582.

[19] Hauser, Ronald G., et al. "A Machine Learning Model to Successfully Predict Future Diagnosis of Chronic Myelogenous Leukemia With Retrospective Electronic Health Records Data." American journal of clinical pathology 156.6 (2021): 1142-1148.

[20] Ali, Farman, et al. "A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion." Information Fusion 63 (2020): 208-222.

[21] Dong, Hongbin, et al. "A novel hybrid genetic algorithm with granular information for feature selection and optimization." Applied Soft Computing 65 (2018): 33-46.

[22] Hans, Rahul, Harjot Kaur, and Navreet Kaur. "Opposition-based Harris Hawks optimization algorithm for feature selection in breast mass classification." Journal of Interdisciplinary Mathematics 23.1 (2020): 97-106.

[23] https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset